Medidas de associação

Regressão Linear Simples

Felipe Figueiredo

UNIAN - Centro Universitário Anhanguera de Niterói

Sumário

Sumário

Modelos estatísticos

Modelos servem para:

- representar de forma simplificada fenômenos, experimentos, dados, etc;
- possibilitar análise em cenários controlados, menos complexos que a realidade;
- extrapolar resultados e conclusões.

Modelos estatísticos

Modelos servem para:

- representar de forma simplificada fenômenos, experimentos, dados, etc;
- possibilitar análise em cenários controlados, menos complexos que a realidade;
- extrapolar resultados e conclusões.

Modelos estatísticos

Modelos servem para:

- representar de forma simplificada fenômenos, experimentos, dados, etc;
- possibilitar análise em cenários controlados, menos complexos que a realidade;
- extrapolar resultados e conclusões.

Modelos estatísticos

Ao ajustar um modelo aos dados, podemos:

- fazer predições dentro do intervalo observado para dados que não foram obtidos (interpolação)
- fazer predições fora do intervalo observado (extrapolação)

Modelos estatísticos

Ao ajustar um modelo aos dados, podemos:

- fazer predições dentro do intervalo observado para dados que não foram obtidos (interpolação)
- fazer predições fora do intervalo observado (extrapolação)

Enquanto isso, no mundo real...

ENDEAVOR BRASIL

Como a Netflix sabia que "House Of Cards" seria um sucesso antes mesmo de lançar a série

"House Of Cards" é o assunto do momento e, pasmem, isso não é novidade para a Netflix. A empresa, baseada em sua cultura de resultados, conseguiu determinar o sucesso da série, antes mesmo de ela ser lançada. Entenda como

Seguir + **Endeavor Brasil,** 31 de março de 2016

Compartilhar 1,1 mil

Fonte: `www.administradores.com.br`

Enquanto isso, no mundo real...

Anhanguera

Da matéria

A Netflix analisou seu mercado inteiro para entender qual série iria repercutir melhor. E não foram apenas pesquisas de comportamento. (. . .)

A organização analisou cada clique, pausa, tempo de retenção nas séries e filmes, aceleração ou desaceleração de frames, entre mil outros fatores, até chegar a uma conclusão.

Fonte: `www.administradores.com.br`

Enquanto isso, no mundo real...

Anhanguera

Da matéria

[Eles entenderam] que seu público gostava de séries políticas (. . .)

A empresa também começou a olhar para uma antiga série britânica que fez muito sucesso e estava a venda, com atores e diretores populares dentre seu público alvo.

Fonte: `www.administradores.com.br`

Enquanto isso, no mundo real...

Anhanguera

Da matéria

Se eles tivessem parado de analisar logo no primeiro fator, muito provavelmente a análise teria sido comprometida. (. . .)

Afinal, quantas foram as vezes que não clicamos em uma série, mas não ficamos nela por mais de 5 minutos?

Fonte: `www.administradores.com.br`

Reta de regressão

Definition

Uma reta de regressão (também chamada de reta de melhor ajuste) é a reta para a qual a soma dos erros quadráticos dos resíduos é o mínimo.

- É a reta que melhor se ajusta aos dados
- Minimiza os resíduos

Reta de regressão

Definition

Uma reta de regressão (também chamada de reta de melhor ajuste) é a reta para a qual a soma dos erros quadráticos dos resíduos é o mínimo.

- É a reta que melhor se ajusta aos dados
- Minimiza os resíduos

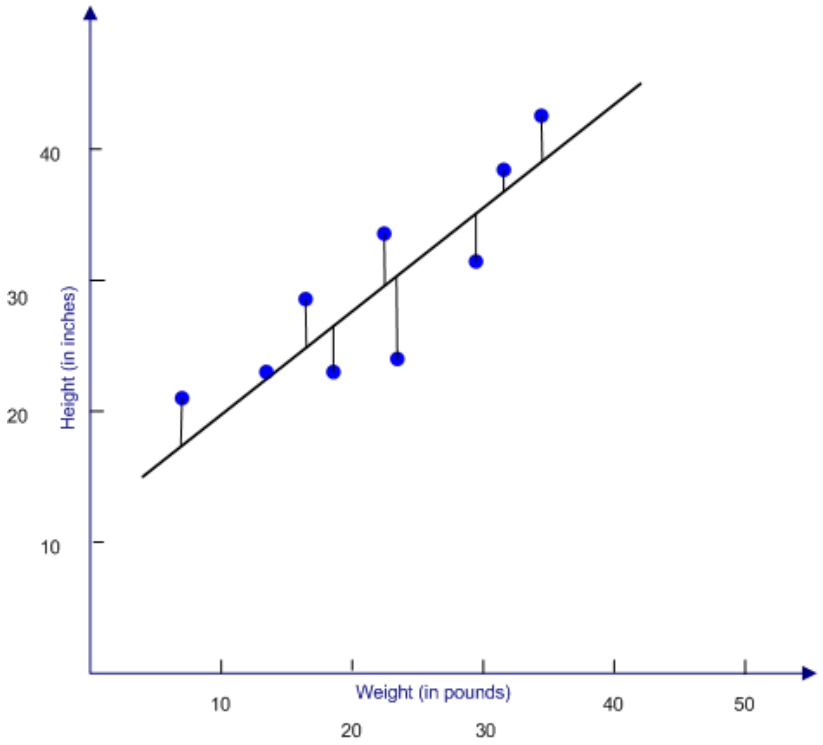
Reta de regressão

Definition

Uma reta de regressão (também chamada de reta de melhor ajuste) é a reta para a qual a soma dos erros quadráticos dos resíduos é o mínimo.

- É a reta que melhor se ajusta aos dados
- Minimiza os resíduos

Resíduos



Definition

Resíduos são a distância entre o dado observado e a reta estimada (modelo).

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula

$$\hat{y} = ax + b$$

- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula

$$\hat{y} = ax + b$$

- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula

$$\hat{y} = ax + b$$

- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula

$$\hat{y} = ax + b$$

- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula

$$\hat{y} = ax + b$$

- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula

$$\hat{y} = ax + b$$

- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula

$$\hat{y} = ax + b$$

- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- . . . e algumas operações entre estes termos

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- . . . e algumas operações entre estes termos

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- . . . e algumas operações entre estes termos

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- . . . e algumas operações entre estes termos

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- . . . e algumas operações entre estes termos

Exercício

Exercício

Dados de gastos com propaganda (x) e vendas (x), ambos em $1000 de uma empresa.

x	2.4	1.6	2.0	2.6	1.4	1.6	2.0	2.2
y	225	184	220	240	180	184	186	215

Qual é a *previsão* de retorno em vendas, para os seguintes gastos com propagandas?

1. 1.5
2. 1.8
3. 2.5

Fonte: Larson & Farber.

Exercício

Exercício

Dados de gastos com propaganda (x) e vendas (x), ambos em $1000 de uma empresa.

x	2.4	1.6	2.0	2.6	1.4	1.6	2.0	2.2
y	225	184	220	240	180	184	186	215

Qual é a *previsão* de retorno em vendas, para os seguintes gastos com propagandas?

1. 1.5
2. 1.8
3. 2.5

Fonte: Larson & Farber.

Exercício

Exercício

Dados de gastos com propaganda (x) e vendas (x), ambos em $1000 de uma empresa.

x	2.4	1.6	2.0	2.6	1.4	1.6	2.0	2.2
y	225	184	220	240	180	184	186	215

Qual é a *previsão* de retorno em vendas, para os seguintes gastos com propagandas?

1. 1.5
2. 1.8
3. 2.5

Fonte: Larson & Farber.

Exercício

Exercício

Dados de gastos com propaganda (x) e vendas (x), ambos em $1000 de uma empresa.

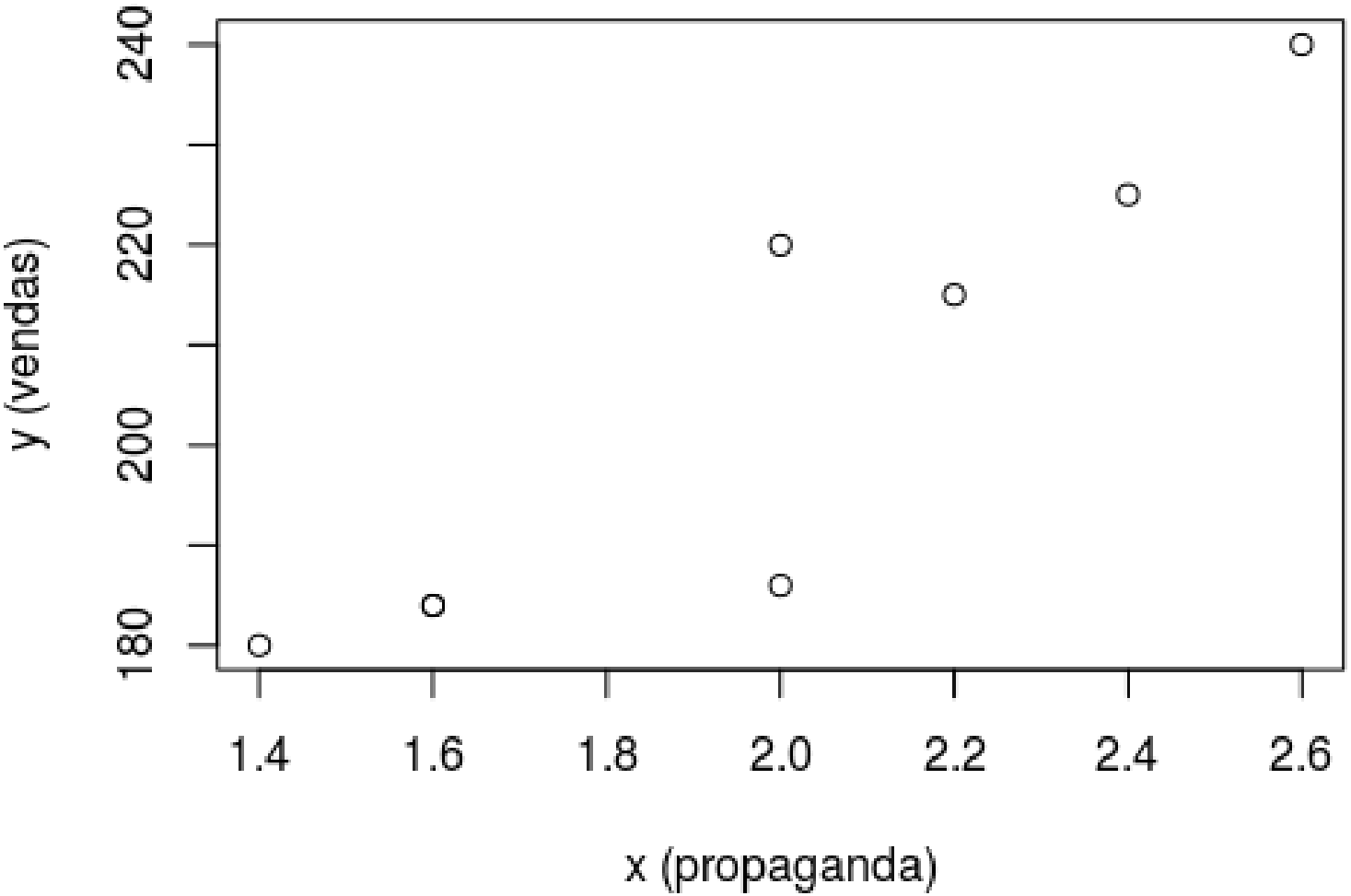
x	2.4	1.6	2.0	2.6	1.4	1.6	2.0	2.2
y	225	184	220	240	180	184	186	215

Qual é a *previsão* de retorno em vendas, para os seguintes gastos com propagandas?

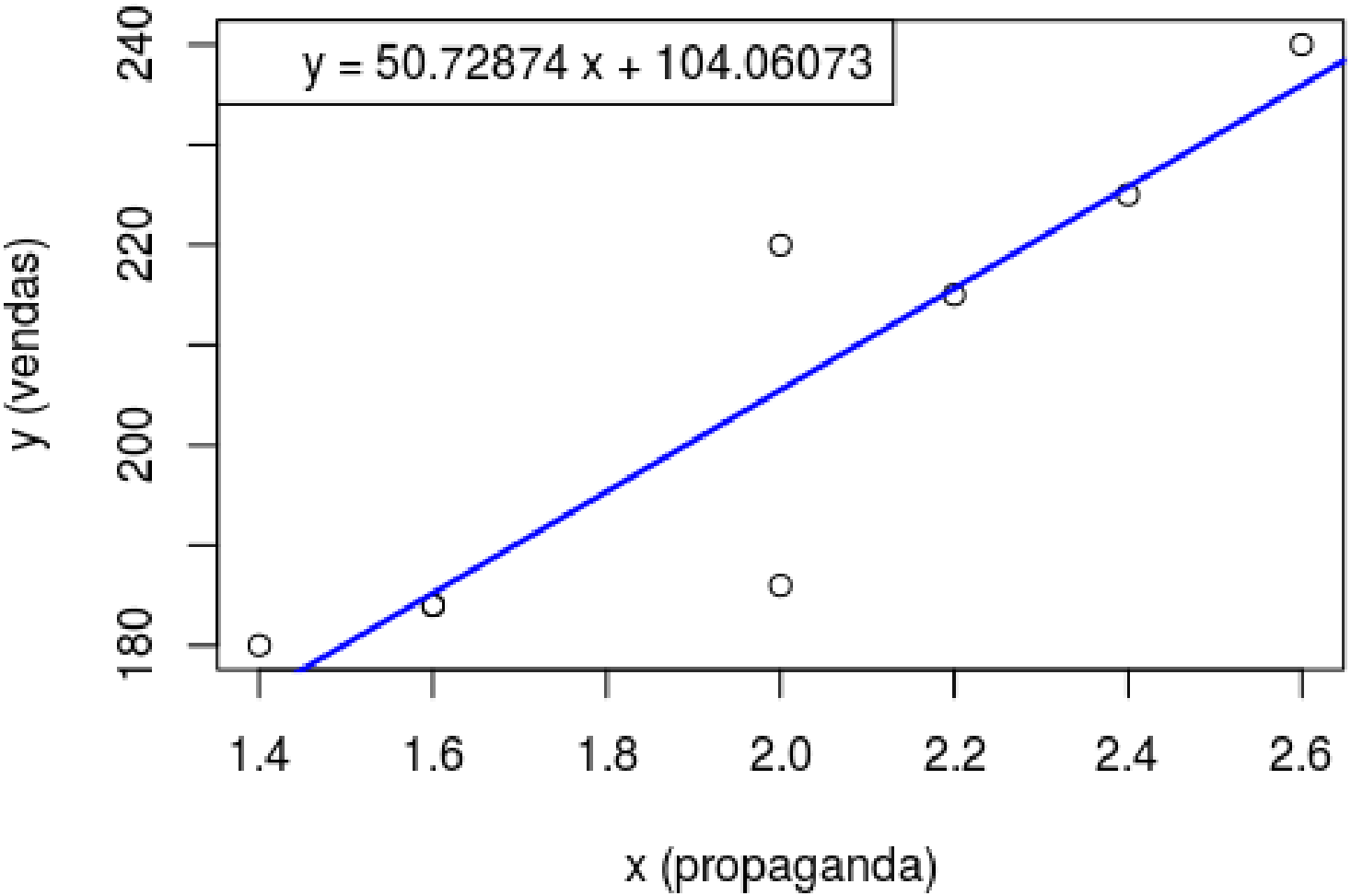
1. 1.5
2. 1.8
3. 2.5

Fonte: Larson & Farber.

Exercício



Exercício


y = 50.72874 x + 104.06073
y (vendas)
180
200
220
240
1.4
1.6
1.8
2.0
2.2
2.4
2.6
x (propaganda)

Exercício

Cola

- $y = 50.72874x + 104.06073$
- $x_1 = 1.5$
- $x_2 = 1.8$
- $x_3 = 2.5$

Solução

1. $y_1 = 50.72874x_1 + 104.06073 = 50.72874 \times 1.5 + 104.06073 = 180.1538 \approx 180.2$
2. $y_2 = 50.72874x_2 + 104.06073 = 50.72874 \times 1.8 + 104.06073 = 195.3725 \approx 195.4$
3. $y_3 = 50.72874x_3 + 104.06073 = 50.72874 \times 2.5 + 104.06073 = 230.8826 \approx 230.9$

Exercício

Cola

- $y = 50.72874x + 104.06073$
- $x_1 = 1.5$
- $x_2 = 1.8$
- $x_3 = 2.5$

Solução

1. $y_1 = 50.72874x_1 + 104.06073 = 50.72874 \times 1.5 + 104.06073 = 180.1538 \approx 180.2$
2. $y_2 = 50.72874x_2 + 104.06073 = 50.72874 \times 1.8 + 104.06073 = 195.3725 \approx 195.4$
3. $y_3 = 50.72874x_3 + 104.06073 = 50.72874 \times 2.5 + 104.06073 = 230.8826 \approx 230.9$

Exercício

Cola

- $y = 50.72874x + 104.06073$
- $x_1 = 1.5$
- $x_2 = 1.8$
- $x_3 = 2.5$

Solução

1. $y_1 = 50.72874x_1 + 104.06073 = 50.72874 \times 1.5 + 104.06073 = 180.1538 \approx 180.2$
2. $y_2 = 50.72874x_2 + 104.06073 = 50.72874 \times 1.8 + 104.06073 = 195.3725 \approx 195.4$
3. $y_3 = 50.72874x_3 + 104.06073 = 50.72874 \times 2.5 + 104.06073 = 230.8826 \approx 230.9$

Exercício

Cola

- $y = 50.72874x + 104.06073$
- $x_1 = 1.5$
- $x_2 = 1.8$
- $x_3 = 2.5$

Solução

1. $y_1 = 50.72874x_1 + 104.06073 = 50.72874 \times 1.5 + 104.06073 = 180.1538 \approx 180.2$
2. $y_2 = 50.72874x_2 + 104.06073 = 50.72874 \times 1.8 + 104.06073 = 195.3725 \approx 195.4$
3. $y_3 = 50.72874x_3 + 104.06073 = 50.72874 \times 2.5 + 104.06073 = 230.8826 \approx 230.9$

Exercício

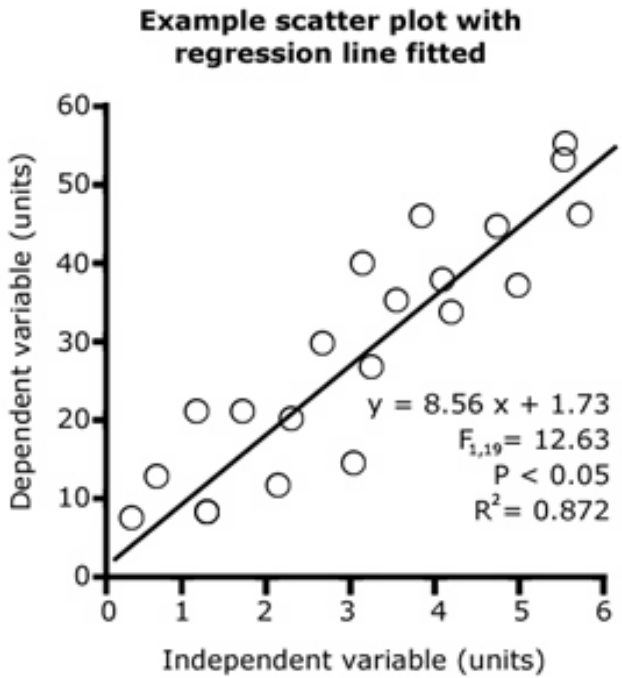
Cola

- $y = 50.72874x + 104.06073$
- $x_1 = 1.5$
- $x_2 = 1.8$
- $x_3 = 2.5$

Solução

1. $y_1 = 50.72874x_1 + 104.06073 =$
   $50.72874 \times 1.5 + 104.06073 = 180.1538 \approx 180.2$
2. $y_2 = 50.72874x_2 + 104.06073 =$
   $50.72874 \times 1.8 + 104.06073 = 195.3725 \approx 195.4$
3. $y_3 = 50.72874x_3 + 104.06073 =$
   $50.72874 \times 2.5 + 104.06073 = 230.8826 \approx 230.9$

Análise de Regressão



- A qualidade do ajuste do modelo de regressão é determinado pelo coeficiente de determinação r^2

Sumário

Coeficiente de Determinação r^2

Definition

O coeficiente de determinação r^2 é a relação da variação explicada com a variação total.

$$r^2 = \frac{\text{variação explicada}}{\text{variação total}}$$

- Lembrando: r^2 é o quadrado de r!

Coeficiente de Determinação r^2

Definition

O coeficiente de determinação r^2 é a relação da variação explicada com a variação total.

$$r^2 = \frac{\text{variação explicada}}{\text{variação total}}$$

- Lembrando: r^2 é o quadrado de r!

Coeficiente de Determinação r^2

- Qual é a porcentagem da variação dos dados pode ser explicada pela reta regressora?
- O coeficiente r^2 é a fração da variância que é compartilhada entre X e Y.
- Como r está sempre entre -1 e 1, r^2 está sempre entre 0 e 1.

Coeficiente de Determinação r^2

- Qual é a porcentagem da variação dos dados pode ser explicada pela reta regressora?
- O coeficiente r^2 é a fração da variância que é compartilhada entre X e Y.
- Como r está sempre entre -1 e 1, r^2 está sempre entre 0 e 1.

Coeficiente de Determinação r^2

- Qual é a porcentagem da variação dos dados pode ser explicada pela reta regressora?
- O coeficiente r^2 é a fração da variância que é compartilhada entre X e Y.
- Como r está sempre entre -1 e 1, r^2 está sempre entre 0 e 1.

Coeficiente de Determinação r^2

- Além disso, $r^2 \leq |r|$
- Por que?

Compare os seguintes números entre 0 e 1:

$$\frac{1}{2} \text{ e } \left(\frac{1}{2}\right)^2 = \frac{1}{4} \Rightarrow \frac{1}{4} \leq \frac{1}{2}$$

$$\frac{1}{3} \text{ e } \left(\frac{1}{3}\right)^2 = \frac{1}{9} \Rightarrow \frac{1}{9} \leq \frac{1}{3}$$

Coeficiente de Determinação r^2

- Além disso, $r^2 \leq |r|$
- Por que?

Compare os seguintes números entre 0 e 1:

$$\frac{1}{2} \text{ e } \left(\frac{1}{2}\right)^2 = \frac{1}{4} \Rightarrow \frac{1}{4} \leq \frac{1}{2}$$

$$\frac{1}{3} \text{ e } \left(\frac{1}{3}\right)^2 = \frac{1}{9} \Rightarrow \frac{1}{9} \leq \frac{1}{3}$$

Coeficiente de Determinação r^2

- Além disso, $r^2 \leq |r|$
- Por que?

Compare os seguintes números entre 0 e 1:

$$\frac{1}{2} \text{ e } \left(\frac{1}{2}\right)^2 = \frac{1}{4} \Rightarrow \frac{1}{4} \leq \frac{1}{2}$$

$$\frac{1}{3} \text{ e } \left(\frac{1}{3}\right)^2 = \frac{1}{9} \Rightarrow \frac{1}{9} \leq \frac{1}{3}$$

Interpretação

- Se a correlação é 0, então X e Y não variam juntos (independentes)
- Se a correlação é positiva, então quando uma aumenta, a outra aumenta em proporção direta (linear)
- Se a correlação é negativa, então quando uma aumenta, a outra diminui em proporção inversa (linear)

Interpretação

- Se a correlação é 0, então X e Y não variam juntos (independentes)
- Se a correlação é positiva, então quando uma aumenta, a outra aumenta em proporção direta (linear)
- Se a correlação é negativa, então quando uma aumenta, a outra diminui em proporção inversa (linear)

Interpretação

- Se a correlação é 0, então X e Y não variam juntos (independentes)
- Se a correlação é positiva, então quando uma aumenta, a outra aumenta em proporção direta (linear)
- Se a correlação é negativa, então quando uma aumenta, a outra diminui em proporção inversa (linear)

Cuidado!

- Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável
- Ex: em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita
- Mas comprar mais telefones não vai salvar crianças!
- Explicação alternativa: a melhoria da condições financeiras pode afetar ambas as variáveis

Cuidado!

- Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável
- Ex: em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita
- Mas comprar mais telefones não vai salvar crianças!
- Explicação alternativa: a melhoria da condições financeiras pode afetar ambas as variáveis

Cuidado!

- Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável
- Ex: em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita
- Mas comprar mais telefones não vai salvar crianças!
- Explicação alternativa: a melhoria da condições financeiras pode afetar ambas as variáveis

Cuidado!

- Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável
- Ex: em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita
- Mas comprar mais telefones não vai salvar crianças!
- Explicação alternativa: a melhoria da condições financeiras pode afetar ambas as variáveis

Causa x efeito

- Se há uma relação de causalidade entre as duas variáveis, a correlação será não nula (positiva ou negativa)
- Quanto maior for a relação de dependência entre as variáveis, maior será o módulo da correlação.
- Se as variáveis não são relacionadas, a correlação será nula.

Causa x efeito

- Se há uma relação de causalidade entre as duas variáveis, a correlação será não nula (positiva ou negativa)
- Quanto maior for a relação de dependência entre as variáveis, maior será o módulo da correlação.
- Se as variáveis não são relacionadas, a correlação será nula.

Causa x efeito

- Se há uma relação de causalidade entre as duas variáveis, a correlação será não nula (positiva ou negativa)
- Quanto maior for a relação de dependência entre as variáveis, maior será o módulo da correlação.
- Se as variáveis não são relacionadas, a correlação será nula.

Causalidade?

- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Causalidade?

- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Causalidade?

- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Causalidade?

- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Causalidade?

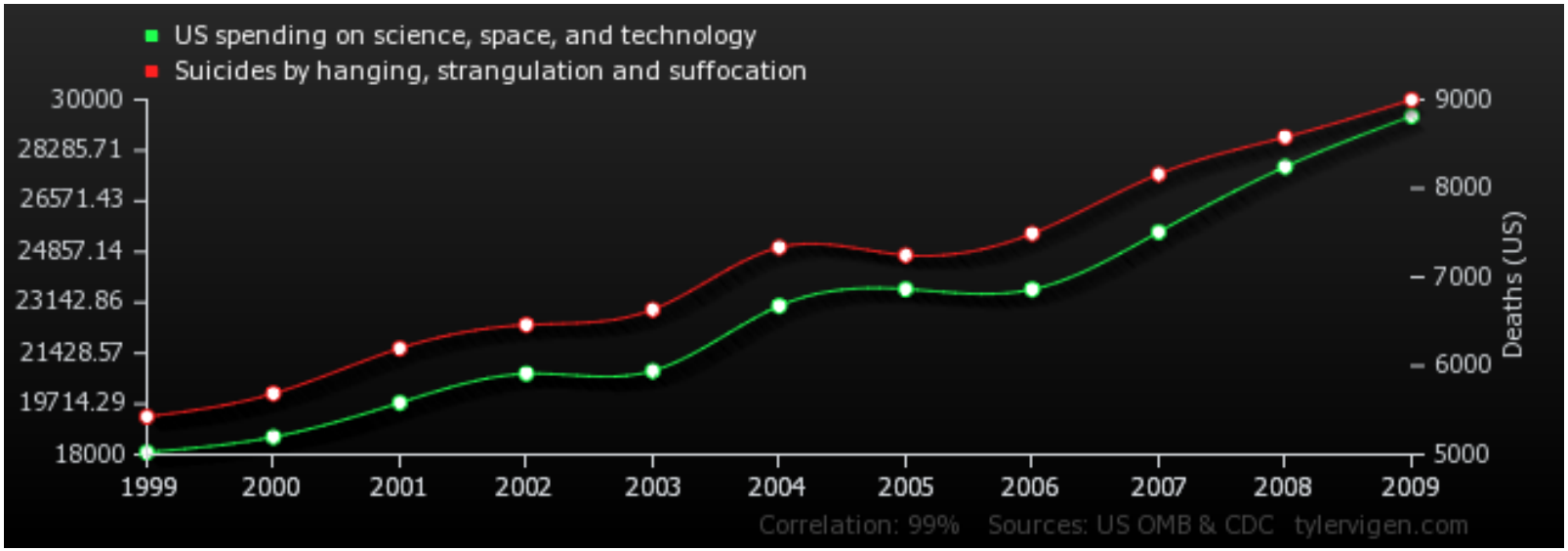
- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Exemplo

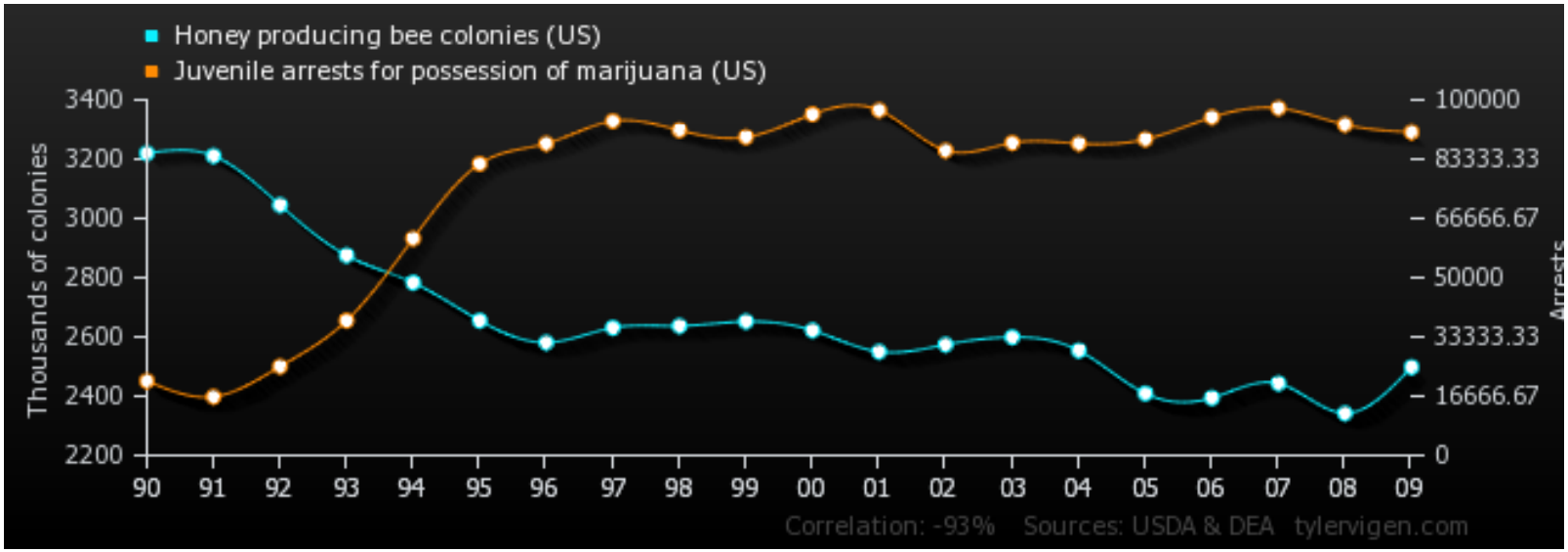
Gasto com C&T (EUA) x Suicídios por enforcamento



Correlação: 0.992082
(Fonte: Spurious correlations)

Exemplo

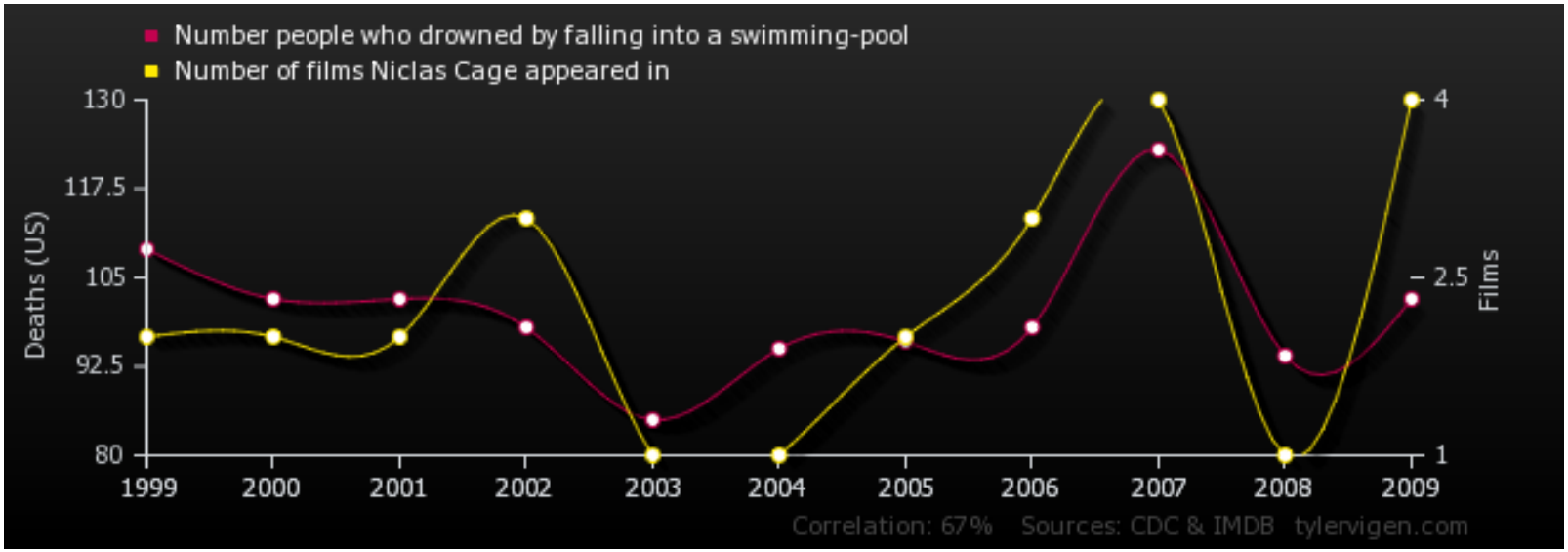
Produção de mel x Prisões por posse de maconha



Correlação: -0.933389
(Fonte: Spurious correlations)

Exemplo

Afogamentos em piscina x Filmes com Nicholas Cage



Correlação: 0.666004
(Fonte: Spurious correlations)

Causa e efeito

Ao encontrar uma forte correlação, deve-se sempre se perguntar:

1. Há uma relação direta de causa e efeito entre as variáveis? (X causa Y?)
2. Há uma relação inversa de causa e efeito entre as variáveis? (Y causa X?)
3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável (ou mais) que não foi analisada?
4. É possível que a relação entre duas variáveis seja uma coincidência?

Causa e efeito

Ao encontrar uma forte correlação, deve-se sempre se perguntar:

1. Há uma relação direta de causa e efeito entre as variáveis? (X causa Y?)
2. Há uma relação inversa de causa e efeito entre as variáveis? (Y causa X?)
3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável (ou mais) que não foi analisada?
4. É possível que a relação entre duas variáveis seja uma coincidência?

Causa e efeito

Ao encontrar uma forte correlação, deve-se sempre se perguntar:

1. Há uma relação direta de causa e efeito entre as variáveis? (X causa Y?)
2. Há uma relação inversa de causa e efeito entre as variáveis? (Y causa X?)
3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável (ou mais) que não foi analisada?
4. É possível que a relação entre duas variáveis seja uma coincidência?

Causa e efeito

Anhanguera

Ao encontrar uma forte correlação, deve-se sempre se perguntar:

1. Há uma relação direta de causa e efeito entre as variáveis? (X causa Y?)
2. Há uma relação inversa de causa e efeito entre as variáveis? (Y causa X?)
3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável (ou mais) que não foi analisada?
4. É possível que a relação entre duas variáveis seja uma coincidência?

Resumo

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
- Qual proporção (porcentagem) da variabilidade pode ser explicada pelas variáveis analisadas?
- Quão bem a reta regressora se ajusta aos dados?

Resumo

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
- Qual proporção (porcentagem) da variabilidade pode ser explicada pelas variáveis analisadas?
- Quão bem a reta regressora se ajusta aos dados?

Resumo

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
- Qual proporção (porcentagem) da variabilidade pode ser explicada pelas variáveis analisadas?
- Quão bem a reta regressora se ajusta aos dados?

Resumo

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
- Qual proporção (porcentagem) da variabilidade pode ser explicada pelas variáveis analisadas?
- Quão bem a reta regressora se ajusta aos dados?